Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/12BE13Zivolic

RUDARENJE PODATAKA I NAPREDNE ANALITIČKE TEORIJE I METODE

DATA MINING AND ADVANCED ANALYTICAL THEORY AND METHODS

Smiljana Živolić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu prikazan je kratak pregled Poslovne analitike sa akcentom na teorije i metode Rudarenja podataka. Predstavljena je osnovna istraživačka analiza podataka, Klaster analiza kao i Linearna i Logistička regresija.*

Ključne reči: *Rudarenje podataka, poslovna analitika, algoritmi, statistika*

Abstract – *This paper presents a brief overview of Business Analytics with an emphasis on theories and methods of Data Mining. Basic research data analysis, Cluster analysis as well as Linear and Logistic regression are presented.*

Keywords: *Data Mining, Business Analitics, Algorithms, Statistics*

1. UVOD

Podaci su takoreći novo zlato i njegovo iskopavanje radi stvaranja poslovne vrednosti u današnjem kontekstu izuzetno umreženog i digitalnog društva zahteva skup veština koji tradicionalno nismo primenjivali u poslu ili u statistici ili čak u inženjerskim programima.

Ideja da senzorima možemo da povežemo fizičke objekte kao što su domovi, automobili, putevi, čak i kante za smeće i ulična svetla, sa digitalno optimizovanim sistemom upravljanja ide ruku pod ruku sa većim podacima i potrebom za dubljim analitičkim mogućnostima.

Nismo daleko od pametnog frižidera koji će osetiti da vam nedostaje, recimo, jaja, popuniti listu za kupovinu mobilne aplikacije vaše prehrambene prodavnice i organizovati Task Rabbit-a da vam pripremi trgovinu namirnicama. Ili frižider koji pregovara o dogovoru sa vozačem Ubera da vam isporuči večernji obrok. Niti smo daleko od senzora ugrađenih u puteve i vozila koji mogu izračunati zagušenost saobraćaja, pratiti istrošenost kolovoza, evidentirati upotrebu vozila i faktorirati ih na dinamičke cene zasnovane na upotrebi, stope osiguranja, pa čak i oporezivanje. Ovaj hrabri novi svet podstaknuće analitika i sposobnost da se podaci iskoriste za konkurentsku prednost.

Poslovna analitika je disciplina u nastajanju koja će zasigurno pomoći da idemo u korak sa ovim novim talasom.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kupusinac, vanr. prof.

Ovaj rad usmeren je na Poslovnu analitiku, njene delove kao i savremenu primenu u različitim sferama današnjice.

2. POSLOVNA ANALITIKA

Poslovna analitika je proces prikupljanja, sortiranja, obrade i proučavanja poslovnih podataka i korišćenje statističkih modela i iterativnih metodologija za korišćenje podataka u poslovne uvide.

Cilj poslovne analitike jeste da utvrdi koji su skupovi podataka korisni i kako se mogu iskoristiti za rešavanje problema i povećanje efikasnosti, produktivnosti i prihoda poslovanja.

Kao podskup poslovne inteligencije (*eng. Business Inteligence*), poslovna analitika se generalno primenjuje s ciljem identifikovanja podataka koji se mogu primeniti u određenu svrhu.

Poslovna inteligencija je obično opisna, usredsređena je na strategije i alate koji se koriste za prikupljanje, identifikovanje i kategorizaciju sirovih podataka i izveštavanje o prošlim ili trenutnim događajima.

Poslovna analitika je više propisana, posvećena je metodologiji pomoću koje se podaci mogu analizirati, prepoznati obrasci i modeli razvijati kako bi se razjasnili prošli događaji, stvorile predviđanja za buduće događaje i preporučile akcije za maksimalizovanje idealnih ishoda.

Sofisticirani podaci, kvantitativna analiza i matematički modeli koriste se od strane poslovnih analitičara da bi dizajnirali rešenja za probleme vođene podacima. Oni mogu da koriste statistiku, informacione sisteme, računarsku nauku i operativna istraživanja kako bi proširili svoje razumevanje složenih skupova podataka i veštačke inteligencije, dubokog učenja (*eng. Deep Learning*) i neuronskih mreža kako bi mikrosegmentisali dostupne podatke i identifikovali obrasce.

Ove informacije se zatim mogu iskoristiti za tačno predviđanje budućih događaja povezanih sa delovanjem potrošača ili tržišnim trendovima i za preporučivanje koraka koji potrošače mogu usmeriti ka željenom cilju.

Poslovna analitika se može posmatrati kroz niz kategorija i komponenti koje svako analitičko rešenje ima. U teoriji se pojavljuje osam osnovnih kategorija:

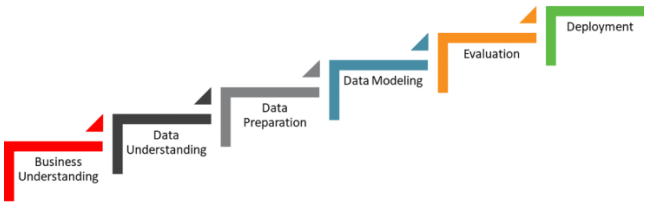
- Agregacija podataka
- Rudarenje podataka
- Asocijacija I identifikacija sekvence
- Rudarenje teksta
- Predviđanje
- Prediktivna analitika
- Vizualizacija podataka

3. RUDARENJE PODATAKA

Šta je Data Mining? Kakva je veza između Big Data i Data Mining-a? Zašto je rudarenje podataka danas veoma važno? To su sve pitanja sa kojima se danas sve češće susrećemo. Odgovori na njih biće dati u ovom i u narednim poglavljima.

Data Mining je interdisciplinarno polje koje zahteva znanje iz mašinskog učenja, veštačke inteligencije i matematičke statistike za pronalaženje i izdvajanje obrazaca iz skupova podataka koji prevazilazi mogućnosti SQL jezika.

Rudarenje podataja bavi se otkrivanjem znanja i pronalaženjem obrazaca u skupovima podataka kroz proces primene modela na podatke . Model, srž procesa rudarstva podataka, je tehnika i algoritmi koji se primenjuju na podatke za pronalaženje sličnosti, obrazaca i grupisanje podataka. Uobičajeni modeli pretraživanja podataka su klasifikacija, pravila pridruživanja i klasterizacija. U ovom istraživanju fokus je na grupisanju i pronalaženju čestih skupova predmeta. Rudarstvo podataka ima širok spektar primena u nauci i inženjerstvu. Na primer, u prognozi vremena, model klasifikacije se može koristiti za predviđanje vremena za sledeći dan na osnovu prethodnih podataka. Drugi primer je za predloge filmova u sistemima koji preporučuju.Na osnovu nekih korisničkih preferencija, mogu se predvideti drugi filmovi koje korisnik preferira. Algoritmi klaster modela takođe imaju širok spektar aplikacija kao što su dovršavanje scene, sažimanje podataka i tehnike oporavka podataka. Na kraju, česti set predmeta se široko koristi u maloprodajnim sistemima za predviđanje kupaca u kupovnim navikama, kao i u dizajnu kataloga.



Slika 1. Šest koraka Rudarenja podataka [1]

3.1. Izazovi u Rudarenju podataka

Iako je moćan proces, rudarenje podataka ometaju sve veća količina i složenost velikih podataka (*eng. Big data*). Tamo gde firme svakodnevno prikupljaju egzabajte podataka, donosioci odluka trebaju načine za izdvajanje, analizu i sticanje uvida iz svog obilnog spremišta podataka.

Izazovi velikih podataka su plodni i prodiru u svako polje koje prikuplja, skladišti i analizira podatke. Veliki podaci karakterišu četiri glavna izazova: obim, raznolikost, istinitost i brzina. Cilj rudarenja podacima je da posreduje u ovim izazovima i otkrije vrednost podataka.

Kako brzina podataka nastavlja da povećava obim i raznolikost podataka, preduzeća moraju da skaliraju ove modele i primenjuju ih u celoj organizaciji. Omogućavanje punih prednosti rudarenja podataka ovim modelima zahteva značajna ulaganja u računarsku infrastrukturu i procesorsku snagu. Da bi dostigle razmere, organizacije moraju da kupe i održavaju moćne računare, servere i softver dizajnirane da obrađuju veliku količinu i raznolikost podataka kompanije.

Povećani zahtevi za skladištenjem podataka primorali su mnoge kompanije da se okrenu računarstvu i cloud skladištenju. Iako je cloud osnažio mnoga savremena dostignuća u rudarstvu podataka, priroda usluge stvara značajne pretnje privatnosti i bezbednosti. Organizacije moraju zaštititi svoje podatke od zlonamernih podataka da bi održale poverenje svojih partnera i kupaca.

3.2. Vrste Rudarenja podataka

Rudarenje podataka ima dva osnovna procesa: učenje pod nadzorom (nadgledano) i bez nadzora (nenadgledano).

Nadgledno učenje

Cilj učenja pod nadzorom je predviđanje ili klasifikacija. Najlakši način da se konceptualizuje ovaj proces je traženje jedne izlazne promenljive. Proces se smatra učenjem pod nadzorom ako je cilj modela predviđanje vrednosti posmatranja.

Jedan od primera su filteri za neželjenu poštu, koji koriste nadzirano učenje za klasifikaciju dolaznih e-adresa kao neželjenog sadržaja i automatsko uklanjanje ovih poruka iz prijemnog sandučeta.

Uobičajeni analitički modeli koji se koriste u pristupima nadgledanog rudarenja podataka su:

Regresije

- ***Vremenske serije*** - Modeli vremenskih serija su alati za predviđanje koji koriste vreme kao primarnu nezavisnu promenljivu. Trgovci na malo, kao što je Maci's, primenjuju modele vremenskih serija da bi predvideli potražnju za proizvodima u funkciji vremena i koristili predviđanje za tačno planiranje i skladištenje zaliha sa potrebnim nivoom zaliha.
- ***Klasifikacija*** - Klasifikaciono drveće je tehnika prediktivnog modeliranja koja se može koristiti za predviđanje vrednosti i kategoričkih i kontinuiranih ciljnih promenljivih. Na osnovu podataka, model će stvoriti skupove binarnih pravila za razdvajanje i grupisanje najvećeg udela sličnih ciljnih promenljivih. Poštujući ta pravila, grupa u koju spada novo zapažanje postaće njena predviđena vrednost.
- ***Neuronske mreže*** - Neuronske mreže koriste ulaze i, na osnovu njihove veličine, „pucaće" ili „neće aktivirati" svoj čvor na osnovu zahteva praga. Ovaj signal ili njegov nedostatak se zatim kombinuje sa ostalim „ispaljenim" signalima u skrivenim slojevima mreže, gde se proces ponavlja sve dok se ne kreira izlaz.
- ***K-najbiliži sused***

Nenadgledano učenje

Zadaci bez nadzora se fokusiraju na razumevanje i opisivanje podataka kako bi se otkrili osnovni obrasci unutar njih. Sistemi preporuka koriste nenadgledano učenje kako bi pratili obrasce korisnika i pružali im personalizovane preporuke za poboljšanje korisničkog iskustva.

4. KLASTER ANALIZA

Generalno, klasterisanje je upotreba nenadgledanih tehnika za grupisanje sličnih objekata. U mašinskom učenju, nenadgledano se odnosi na problem pronalaženja skrivene strukture u neobeleženim podacima. Tehnike klasterovanja nisu pod nadzorom u smislu da naučnik unapred ne određuje labele koje primenjuje na klastere. Struktura podataka opisuje objekte od interesa i određuje kako najbolje da se objekti grupišu.

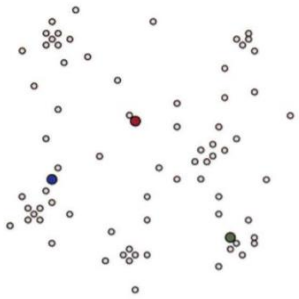
4.1. K-Means

S obzirom na kolekciju predmeta od kojih svaki ima *n* merljivih atributa, K-means je analitička tehnika koja, za izabranu vrednost *k*, identifikuje k klastera objekata na osnovu blizine objekata do centra k grupe. Centar se određuje kao aritmetički prosek (srednja vrednost) n-dimenzionalnog vektora svakog klastera atributa. Ovaj odeljak opisuje algoritam za određivanje k-vrednosti, kao i kako to najbolje primeniti tehniku na nekoliko slučajeva upotrebe.

Grupisanje se često koristi kao uvod u klasifikaciju. Jednom kada su klasteri identifikovani, mogu se primeniti oznake na svakom klasteru da klasifikuje svaku grupu na osnovu njenih karakteristika. Grupisanje je prvenstveno istraživačka tehnika otkrivanja skrivenih struktura podataka, možda kao uvod u fokusiraniju analizu ili procese odlučivanja. Neke specifične primene K-means-a su obrada slike, medicinska i korisnička segmentacija.

K-means algoritam za pronalaženje k klastera može se opisati u sledeća četiri koraka:

1. Izaberemo vrednost *k* i k početno za centroide. U ovom primeru, k = 3, a početni centroidi su označeni tačkama osenčenim crvenom, zelenom, i plavom na *slici 2*.



Slika 2. Korak 1 u pronalaženju K klastera [1]

2. Izračunavamo udaljenost od svake tačke podataka (x,y_i) do svakog centroida. Dodelimo svaku tačku najbližem centroidu. Ovo definiše prvih k klastera.
   U dve dimenzije, rastojanje d između bilo koje dve tačke, (x_1,y_1) i (x_2, y_2), u ravni se tipično izražava korišćenjem Euklidove mere daljine date u jednačini:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

3. Izračunavamo centroid, centar mase svakog novo definisanog skupa iz koraka 2. Na *slici 10*. izračunati centroidi u koraku 3 su svetlo osenčene tačke odgovarajuće boje. U dve dimenzije, centroid (x_c,y_c) od m tačaka u k-means klasteru izračunava se na sledeći način u jednačini :

$$(x_c, y_c) = \left(\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m}\right) \quad (2)$$

- Ponavljamo korake 2 i 3 dok se algoritam ne konvergira u odgovor.
- Dodeljujemo svaku tačku najbližem centoridu izračunatom u koraku 3.
- Izračunavamo centorid novodefinisanih klastera.
- Ponavljamo dok algoritam ne dođe do konačnog odgovora.

Konvergencija se postiže kada se izračunati centroidi ne promene ili kada centroidi i dodeljene tačke osciliraju napred-nazad od jedne do druge iteracije. Može doći do ovog drugog slučaja kada postoji jedna ili više tačaka na jednakim udaljenostima od izračunatog centorida.

5. REGRESIJA

Generalno, regresijska analiza pokušava da objasni uticaj koji skup promenljivih ima na ishod druge promenljive od interesa. Često se promenljiva ishoda naziva zavisnom promenljivom jer ishod zavisi od ostalih promenljivih. Ove dodatne promenljive se ponekad nazivaju ulazne promenljive ili nezavisne promenljive.

Linearna regresija je korisno sredstvo za odgovor na prvo pitanje, a logistička regresija je popularan metod za odgovor na drugo pitanja. Ovo poglavlje ispituje ove dve tehnike regresije i objašnjava kada je jedna tehnika pogodnija od druge.

Regresijska analiza korisno je objašnjenje koje može identifikovati ulazne promenljive koje imaju najveći statistički uticaj na ishod. Sa takvim znanjem i uvidom, promene životne sredine mogu se pokušati proizvesti povoljnije vrednosti ulaznih promenljivih. Na primer, ako se utvrdi da je nivo čitanja desetogodišnjaka odličan prediktor uspeha učenika u srednjoj školi i faktor njihovog pohađanja fakulteta, onda se može razmotriti, primeniti dodatni naglasak na čitanju i ocenjeno radi poboljšanja nivoa čitanja učenika u mlađem uzrastu.

5.1. Linearna regresija

Linearna regresija je analitička tehnika koja se koristi za modeliranje odnosa između nekoliko ulaznih promenljivih i kontinuirane promenljive ishoda. Ključna pretpostavka jeste da je veza između ulazne promenljive i promenljive ishoda linearna. Iako se ova pretpostavka može činiti restriktivnom, često je moguće pravilno transformisati ulazne ili ishodne promenljive kako bi se postigla linearna veza između izmenjenih promenljivih ulaznih i ishodnih.

Model linearne regresije pretpostavlja da postoji linearni odnos između ulaznih promenljivih i promenljive ishoda. Ovaj odnos se može izraziti kao što je prikazano u jednačini:

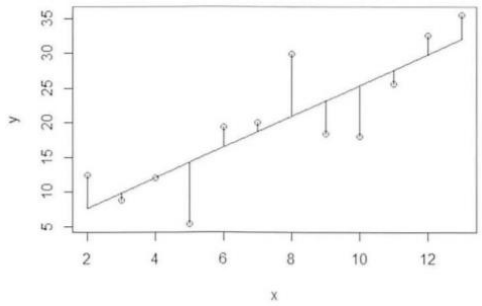
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \xi \quad (3)$$

U linearnom modelu, β_j predstavljaju nepoznate *p* parametre. Procene za ove nepoznate parametre su odabrane tako da u proseku model pruža razumnu procenu prihoda osobe na osnovu starosti i obrazovanja. Drugim rečima, ugrađeni model treba da umanji ukupnu grešku između linearnog modela i stvarnih zapažanja.

Obični najmanji kvadrati su uobičajena tehnika za procenu parametara.
Da bismo ilustrovali kako ova tehnika radi, pretpostavimo da postoji samo jedna ulazna promenljiva, *x*, za izlaznu promenljivu *y*.
Cilj je pronaći liniju koja najbolje aproksimira odnos između izlazne promenljive i ulaznih promenljivih. Cilj najmanjih kvadrata je pronaći liniju kroz ove tačke koja minimalizuje broj kvadrata razlike između svake tačke i prave u vertikalnom smeru. Drugim rečima, pronađemo vrednosti -10 i.-1, tako da je zbir prikazan u sledećoj jednačini minimiziran.

$$\sum_{i=1}^{n} [y_i + (\beta_0 + \beta_1 x_1)]^2 \quad (4)$$

Na *slici 3.* prikazano je n pojedinačnih rastojanja koje treba kvadratiti, a zatim zbrojiti. Vertikala linije predstavljaju rastojanje između svake posmatrane vrednosti *y* i linije $y = \beta_0 + \beta_1 x$.



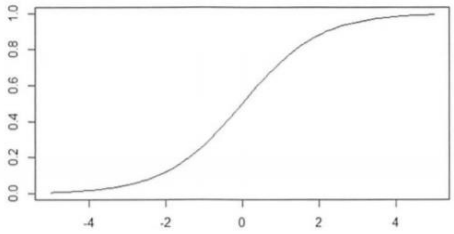
Slika 3.

5.2. Logistička regresija

Logistička regresija se zasniva na logističkoj funkciji f(y) datoj u formuli:

$$f(y) = \frac{e^y}{1+e^y}, \text{ za } -\infty < y < \infty \quad (5)$$

Primetimo da y → ∞, f(y) →1 i y → -∞, f(y) →0, tako da na *slici 4.* vidimo da vrednost logističke funkcije f(y) varira 0 do 1.



Slika 4.

Budući da je opseg f (y) u interval (0, 1), čini se da je logistička funkcija odgovarajuća funkcija za modeliranje verovatnoće da se dogodi određeni ishod. Kako se vrednost y povećava, verovatnoća ishoda se povećava. U bilo kojem predloženom modelu, da bi se predvidela verovatnoća ishoda, y treba da bude funkcija ulaznih promenljivih. U logističkoj regresiji y se izražava kao linearna funkcija ulaznih promenljivih.

6. ZAKLJUČAK

Praksa nauke o podacima može se najbolje opisati kao kombinacija analitičkog inženjerstva i istraživanja. Posao predstavlja problem koji bismo želeli da rešimo. Retko je poslovni problem jedan od naših osnovnih zadataka za rudarenje podacima. Razlažemo problem u podzadatke za koje mislimo da ih možemo rešiti, obično počevši od postojećih alata.

Za neke od ovih zadataka možda ne znamo koliko dobro možemo da ih rešimo, pa moramo da istražimo podatke i izvršimo procenu da bismo to videli. Ako to ne uspe, možda ćemo morati da pokušamo nešto sasvim drugo. U tom procesu možemo otkriti znanje koje će nam pomoći da rešimo problem koji smo zacrtali da rešimo ili ćemo otkriti nešto neočekivano što nas vodi do drugih važnih uspeha.

7. LITERATURA

[1] Galit Shmueli, Peter C. Bruce, Nitin R. Patel, Data Mining for Business Analitics, Third edition
[2] Foster Provost & Tom Fawcett, Data Science for Business
[3] T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review, October 2012.
[4] The R Project for Statistical Computing, "The Comprehensive R Archive Network."
[5] P.-N. Tan, V. Kumar, and M. Steinbach, Introduction to Data Mining, Upper Saddle River, NJ: Person, 2013.

Kratka biografija:

Smiljana Živolić rođena je u Novom Sadu 1993. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Primenjene računarske nauke i informatika odbranila je 2020.god.

kontakt:
smiljana.zivolic@gmail.com